

# A Derivative-Free CoMirror Algorithm

Heinz H. Bauschke\*, Warren L. Hare†, Walaa M. Moursi‡

October 23, 2012

## Abstract

We consider  $\min\{f(x) : g(x) \leq 0, x \in X\}$ , where  $X$  is a compact convex subset of  $\mathbb{R}^m$ , and  $f$  and  $g$  are continuous convex functions defined on an open neighbourhood of  $X$ . We work in the setting of derivative-free optimization, assuming that  $f$  and  $g$  are available through a black-box that provides only function values for a lower- $\mathcal{C}^2$  representation of the functions. We present a derivative-free optimization variant of the  $\varepsilon$ -comirror algorithm [3]. Algorithmic convergence hinges on the ability to accurately approximate subgradients of lower- $\mathcal{C}^2$  functions, which we prove is possible through linear interpolation. We provide convergence analysis that quantifies the difference between the function values of the iterates and the optimal function value. We find that the DFO algorithm we develop has the same convergence result as the original gradient-based algorithm. We present some numerical testing that demonstrate the practical feasibility of the algorithm, and conclude with some directions for further research.

**Keywords:** convex optimization, derivative-free optimization, lower- $\mathcal{C}^2$ , approximate subgradient, Non-Euclidean projected subgradient, Bregman distance.

**2010 Mathematics Subject Classification:** Primary 90C25, 90C56; Secondary 49M30, 65K10.

## 1 Introduction

In this paper we introduce a derivative-free linear interpolation-based method for solving constrained optimization problems of the form

$$(P) : \min\{f(x) : g(x) \leq 0, x \in X\}, \quad (1.1)$$

where  $f$  and  $g$  are continuous convex functions defined on a nonempty open convex subset  $O$  of  $\mathbb{R}^m$ , and where the constraint set  $X$  is a nonempty compact convex subset of  $O$ . We further assume that we have access to the lower- $\mathcal{C}^2$  representations of  $f$  and  $g$  and that the problem is feasible

---

\*Mathematics. Irving K. Barber school, University of British Columbia, Kelowna, B.C. V1V 1V7, Canada. Heinz.Bauschke@ubc.ca.

†Mathematics. Irving K. Barber school, University of British Columbia, Kelowna, B.C. V1V 1V7, Canada. Warren.Hare@ubc.ca.

‡Mathematics. Irving K. Barber school, University of British Columbia, Kelowna, B.C. V1V 1V7, Canada. Walaa.Moursi@ubc.ca.

i.e., there exists some  $x_0 \in X$  such that  $g(x_0) \leq 0$ . The algorithm is based on the  $\varepsilon$ -comirror algorithm presented in [3]. Derivative-free optimization (DFO) is a rapidly growing field of research that explores the minimization of a black-box function when first-order information (derivatives, gradients, or subgradients) is unavailable. While the majority of past work in DFO has focused on unconstrained optimization, several methods have recently been introduced for constrained optimization. In constrained optimization, most of the analysis of DFO methods has been done within the framework of *direct search* and *pattern search* methods. That is, methods that do not attempt to build interpolation (or other such) models of the objective function, but instead use concepts like positive bases to ensure convergence. Such methods can be adapted to constrained optimization through techniques by e.g. projecting search directions onto constraint sets [17, 16], “pulling back” search directions onto manifolds [13, 14], the use of filtering techniques [1], or barrier based penalties [2].

On the other hand, fairly little research has explored approaching constrained optimization via model-based DFO methods. Notable in this area is [23, 24], which extends the UOBYQA [20] to constrained optimization (in an algorithm named CONDOR). This paper provides a novel model-based DFO method for linearly constrained optimization. Our algorithm is designed for constraints defined by a given convex function.

Our algorithm is based on the  $\varepsilon$ -comirror algorithm [3]. The  $\varepsilon$ -comirror algorithm finds its roots in mirror-descent methods [19, 5, 4]. These methods can be viewed as nonlinear projected subgradient methods that use a general distance-like function (the Bregman distance) instead of the usual Euclidean squared distance [4]. The  $\varepsilon$ -comirror algorithm adapts the mirror-descent method to work for convex constrained optimization where the constraint set is provided by a convex function. It requires that the problem is additionally constrained by a convex compact set and that the subgradients (of both the constraint function and the objective function) are bounded over this set.

The algorithm presented here differs from previous research in two other notable ways. First, unlike past model-based DFO method, we do not assume that the objective function is  $\mathcal{C}^2$ ; instead, we work with the broader class of lower- $\mathcal{C}^2$  functions (see definition 2.1). Lower- $\mathcal{C}^2$  functions include convex [22, Theorem 10.33] and  $\mathcal{C}^2$  functions (by definition), as well as fully amenable functions [22, Exercise 10.36] and finite max functions (Example 2.3 below). To work with lower- $\mathcal{C}^2$  functions, we develop a method to approximate subgradients for such functions and analyze it for the derivative-free algorithm. In particular, in Theorem 3.3 we define the approximate subgradient for an arbitrary lower- $\mathcal{C}^2$  function and prove that it satisfies an error bound analogous to the one introduced in [8, Theorem 2.11] for the class of  $\mathcal{C}^1$  functions.

The second major difference from previous DFO research is that we present a convergence result that quantifies the difference between the function values of the iterates and the optimal function value. To the best of our knowledge, this provides the first results of this kind for a multivariable DFO method. It is remarkable that the DFO algorithm we develop has the same convergence result as the original gradient-based algorithm presented in [3]. (A quadratically convergent DFO method is developed in [15], but only for functions defined on  $\mathbb{R}$ . Furthermore, in [18], a superlinearly convergent algorithm is presented.)

The remainder of this paper is organized as follows. Section 2 is a brief introduction to the main building blocks we use. First, we provide the definition of the class of lower- $\mathcal{C}^2$  functions and some properties. Second, we provide the definition of the linear interpolation model of a function  $f$  over a subset  $Y$  of  $\mathbb{R}^m$  and a sufficient condition to be well-defined. Finally, we give

the definition and the main properties of Bregman distances. In Section 3 we give the first key result in Theorem 3.3, on which we build our convergence results. In Section 4 we describe our derivative-free  $\varepsilon$ -comirror algorithm. In Theorem 4.3 we establish the convergence analysis. In Section 5 we provide some numerical results that confirm the practical feasibility of the algorithm. Section 6 summarizes some concluding remarks. To make the presentation self-contained we add Appendix A which includes the proofs of two basic inequalities.

## 2 Auxiliary Results

We shall work in  $\mathbb{R}^m$ , equipped with the usual Euclidean norm  $|\cdot|$ . Throughout the remainder of the paper, we suppose that

$O$  is a nonempty open convex subset of  $\mathbb{R}^m$ .

Recall that for a convex function  $f : O \rightarrow \mathbb{R}$ , the subdifferential  $\partial f$  at a point  $x \in O$  is defined by

$$\partial f(x) = \{v \in \mathbb{R}^m : f(y) \geq f(x) + \langle v, y - x \rangle \text{ for all } y \in O\}. \quad (2.1)$$

We denote the *closed* ball in  $\mathbb{R}^m$  centred at  $x_0$  with radius  $\Delta > 0$  by

$$B(x_0; \Delta) = \{x \in \mathbb{R}^m : |x - x_0| \leq \Delta\},$$

and the set of *natural numbers* by

$$\mathbb{N} = \{1, 2, 3, \dots\}.$$

Given  $r \in \mathbb{N}$ , we abbreviate the *unit simplex* in  $\mathbb{R}^r$  by

$$S_r := \left\{ \lambda \in \mathbb{R}^r : \sum_{i=1}^r \lambda_i = 1, \lambda_i \in [0, 1], i \in \{1, \dots, r\} \right\}.$$

Finally, we shall use  $|L|$  to denote the spectral norm of a matrix  $L \in \mathbb{R}^{m \times m}$ .

### The Class of Lower- $\mathcal{C}^k$ Functions

We next introduce the class of lower- $\mathcal{C}^2$  functions.

**Definition 2.1.** [22, Definition 10.29] A function  $f : O \rightarrow \mathbb{R}$  is said to be a lower- $\mathcal{C}^k$  function at  $\bar{x} \in O$  if there exists a neighbourhood  $V = V(\bar{x}) \subseteq O$  and a representation

$$f(x) = \max_{t \in T} f_t(x) \quad (2.2)$$

in which all functions  $f_t$  are of class  $\mathcal{C}^k$  on  $V$ , the index set  $T := T(\bar{x})$  is a compact topological space, and  $f_t$  and the first  $k$  derivatives of  $f_t$  depend continuously not just on  $x \in V$  but even on  $(t, x) \in T \times V$ . In this case we say that (2.2) provides a lower- $\mathcal{C}^k$  representation of  $f$  at  $x \in O$ . The function  $f$  is said to be lower- $\mathcal{C}^k$  on  $O$  if  $f$  is lower- $\mathcal{C}^k$  at every point  $x \in O$ .

The next Lemma provides details regarding when a convex function is lower- $\mathcal{C}^2$ .

**Lemma 2.2.** [22, Theorem 10.33] *Let  $f : O \rightarrow \mathbb{R}$  be convex. Then  $f$  is lower- $\mathcal{C}^2$  on  $O$ .*

Although the class of lower- $\mathcal{C}^2$  functions includes many convex functions [22, Theorem 10.33], it should be noted that our algorithm will require access to a lower- $\mathcal{C}^2$  representation of the objective and constraint functions. The next example shows that any finite max function is not only lower- $\mathcal{C}^2$ , but also provides a natural lower- $\mathcal{C}^2$  representation.

**Example 2.3.** *Let  $f : O \rightarrow \mathbb{R}$  be defined as  $f = \max \{f_1, \dots, f_n\}$ , where each  $f_i$  is of class  $\mathcal{C}^k$  on  $O$ . Then  $f$  is lower- $\mathcal{C}^k$  on  $O$ . (This is the case where  $T$  is  $\{1, \dots, n\}$  equipped with the discrete topology.)*

The value of working with lower- $\mathcal{C}^2$  functions is seen in Lemma 2.4, which demonstrates how to compute the subdifferential of a lower- $\mathcal{C}^2$  function.

**Lemma 2.4.** *Let  $f : O \rightarrow \mathbb{R}$  be a convex function that has a lower- $\mathcal{C}^2$  representation  $f(x) = \max_{t \in T} f_t(x)$  at  $\bar{x} \in O$  and set  $A(\bar{x}) = \operatorname{argmax}_{t \in T} f_t(\bar{x})$ . Then*

$$\partial f(\bar{x}) = \operatorname{conv} \{ \nabla f_t(\bar{x}) \mid t \in A(\bar{x}) \}.$$

*Proof.* Combine [22, Theorem 10.31] and [22, Proposition 8.12].  $\square$

**Theorem 2.5.** [22, Proposition 10.54] *Let  $f : O \rightarrow \mathbb{R}$  be a lower- $\mathcal{C}^2$  function, and let  $X$  be a nonempty compact subset of  $O$ . Then there exists an open set  $O'$  with  $X \subseteq O' \subseteq O$ , such that  $f$  has a common lower- $\mathcal{C}^2$  representation valid at all points  $x \in O'$ , i.e., there exists a compact topological space  $T$ , and a family of functions  $(f_t)_{t \in T}$  defined on  $O'$  such that*

$$f = \max_{t \in T} f_t \quad \text{on } O', \tag{2.3}$$

*and the functions  $(t, x) \mapsto f(t, x)$ ,  $(t, x) \mapsto \nabla f(t, x)$ , and  $(t, x) \mapsto \nabla^2 f(t, x)$  are continuous on  $T \times O'$ .*

To prove convergence of the algorithm introduced in this paper, we require bounds on the subgradients of the objective and the constraint functions. Lemma 2.6 provides a proof of the existence of this bound.

**Lemma 2.6.** *Let  $f : O \rightarrow \mathbb{R}$  be convex, and let  $X$  be a nonempty compact subset of  $O$ . Then*

$$\sup |\partial f(X)| < +\infty.$$

*Proof.* Since  $f$  is convex, Lemma 2.2 implies that  $f$  is lower- $\mathcal{C}^2$  on  $O$ . Since  $X$  is a nonempty compact subset of  $O$ , Theorem 2.5 guarantees the existence of an open subset  $O'$  with  $X \subseteq O' \subseteq O$  such that  $f$  has a common lower- $\mathcal{C}^2$  representation valid at all points  $x \in O'$ . Let  $f = \max_{t \in T} f_t$  be as stated in Theorem 2.5. The definition of lower- $\mathcal{C}^2$  implies that the mapping  $(t, x) \mapsto |\nabla f_t(x)|$  is continuous on  $T \times O'$ . By the Weierstrass Theorem,  $L := \max_{(t, x) \in T \times X} |\nabla f_t(x)| < \infty$ . Now, let  $x \in X$ , and let  $v \in \partial f(x)$ . Using Lemma 2.4 we know that  $v = \sum_{t \in A(x)} \lambda_t \nabla f_t(x)$  for some  $\lambda \in S^r$  where  $r \in \mathbb{N}$  is the number of elements in  $A(x)$ . Therefore

$$|v| = \left| \sum_{t \in A(x)} \lambda_t \nabla f_t(x) \right| \leq \sum_{t \in A(x)} \lambda_t |\nabla f_t(x)| \leq \sum_{t \in A(x)} \lambda_t L = L,$$

and the proof is complete. (Alternatively, one may consider either the lower semicontinuous hull of  $f$  and apply [21, Theorem 24.7], or use [22, Corollary 12.38] after extending  $\partial f$  to a maximally monotone operator.)  $\square$

**Lemma 2.7.** *Let  $f : O \rightarrow \mathbb{R}$  be a lower- $\mathcal{C}^2$  function, and let  $X$  be a nonempty compact convex subset of  $O$ . Let  $O'$ ,  $T$ , and  $(f_t)_{t \in T}$  be as in Theorem 2.5. Then there exists  $K_f \geq 0$  such that  $\nabla f_t$  is  $K_f$ -Lipschitz on  $O'$  for every  $t \in T$ .*

*Proof.* By Theorem 2.5,  $(t, x) \mapsto \nabla^2 f_t(x)$  is continuous on the compact set  $T \times X$ . Therefore, by the Weierstrass theorem,  $K_f := \max_{(t, x) \in T \times X} \|\nabla^2 f_t(x)\| < +\infty$ . Now apply the Mean Value Theorem [12, Theorem 5.1.12]. □

## The Linear Interpolation Model

In our method we use a derivative-free model-based technique. Therefore, in this section we introduce the definition of the linear interpolation model and related facts.

**Definition 2.8.** *Let  $f : O \rightarrow \mathbb{R}$  be a function, and let  $Y = (y_0, y_1, \dots, y_m) \in \mathbb{R}^{m \times (m+1)}$ . If the matrix*

$$Q = \begin{pmatrix} 1 & y_{0,1} & \dots & y_{0,m} \\ 1 & y_{1,1} & \dots & y_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_{m,1} & \dots & y_{m,m} \end{pmatrix}$$

*is invertible, then  $Y$  is said to be a poised tuple centred at  $y_0$ . Moreover, if  $\{y_0, y_1, \dots, y_m\} \subseteq O$  then  $Y$  is said to be a poised tuple centred at  $y_0$  with respect to  $f$ . In this case the linear system*

$$\begin{pmatrix} 1 & y_{0,1} & \dots & y_{0,m} \\ 1 & y_{1,1} & \dots & y_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_{m,1} & \dots & y_{m,m} \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} f(y_0) \\ f(y_1) \\ \vdots \\ f(y_m) \end{pmatrix}.$$

*has a unique solution  $(\alpha_0, \alpha_1, \dots, \alpha_m) \in \mathbb{R}^{m \times (m+1)}$ , and the Linear Interpolation Model of the function  $f$  over  $Y$  is the unique (well defined) function*

$$F : \mathbb{R}^m \rightarrow \mathbb{R} : x \mapsto \alpha_0 + \sum_{i=1}^n \alpha_i x_i.$$

*Note that in this case  $F$  satisfies the interpolation conditions*

$$F(y_i) = f(y_i), \text{ for every } i \in \{0, 1, \dots, m\}.$$

The following Theorem provides the error bound satisfied by the approximate gradient of the linear interpolation model.

**Theorem 2.9.** [8, Theorem 2.11] *Suppose that  $f : O \rightarrow \mathbb{R}$  is  $\mathcal{C}^2$  function on  $O$ . Let  $y_0 \in O$ . Assume that  $Y = (y_0, y_1, \dots, y_m) \in \mathbb{R}^{m \times (m+1)}$  is a poised tuple of sample points centred at  $y_0$  with respect to  $f$ . Set  $\Delta = \max_{1 \leq i \leq m} |y_i - y_0|$ . Suppose that  $B(y_0; \Delta) \subseteq O$ . Let  $\nabla f$  be  $K_f$  Lipschitz over  $B(y_0; \Delta)$ . Then the gradient of the linear interpolation model  $F$  satisfies an error bound of the form*

$$|\nabla f(y) - \nabla F(y)| \leq K\Delta, \text{ for all } y \in B(y_0; \Delta),$$

where

$$K := K_f(1 + \sqrt{m}|\hat{L}^{-1}|/2), \quad L = L(Y) := \begin{pmatrix} y_1 - y_0 \\ y_2 - y_0 \\ \vdots \\ y_m - y_0 \end{pmatrix}, \quad \text{and } \hat{L} = \hat{L}(Y) := \frac{1}{\Delta}L. \quad (2.4)$$

## The Bregman Distance: Definition and Properties

The last building block used in our analysis is the Bregman distance.

**Definition 2.10.** [6] *Let  $\omega : O \rightarrow \mathbb{R}$  be a convex differentiable function. The corresponding Bregman distance  $D_\omega$  is*

$$D_\omega : O \times O \rightarrow \mathbb{R} : (u, v) \mapsto \omega(u) - \omega(v) - \langle \nabla \omega(v), u - v \rangle. \quad (2.5)$$

**Definition 2.11.** [26, Section 3.5] *Let  $C$  be a nonempty convex subset of  $\mathbb{R}^m$ . Let  $\omega : C \rightarrow \mathbb{R}$ . Then  $\omega$  is said to be strongly convex with convexity parameter  $\alpha > 0$ , if for all  $x, y \in C$ ,  $t \in [0, 1]$  we have*

$$\omega(tx + (1-t)y) \leq t\omega(x) + (1-t)\omega(y) - \frac{\alpha}{2}t(1-t)|x - y|^2.$$

Throughout the next arguments we shall assume that  $\omega$  is a strongly convex and differentiable function on a nonempty convex subset of  $\mathbb{R}^m$ , with a convexity parameter  $\alpha > 0$ . In this paper we shall be interested in Bregman distances that are created from strongly convex functions.

The following result is part of the folklore (and established in much greater generality in e.g., [26, Section 3.5]); for completeness we include the proof.

**Lemma 2.12.** *Let  $\omega : O \rightarrow \mathbb{R}$  be a differentiable function. Let  $X$  be a nonempty subset of  $O$ . Then the following are equivalent:*

- (i)  $\omega(\lambda x + (1-\lambda)y) \leq \lambda\omega(x) + (1-\lambda)\omega(y) - \frac{\alpha}{2}\lambda(1-\lambda)|x - y|^2$  for all  $x, y \in X$  and  $\lambda \in ]0, 1[$ .
- (ii)  $D_\omega(x, y) = \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle \geq \frac{\alpha}{2}|x - y|^2$  for all  $x, y \in X$  and  $\lambda \in ]0, 1[$ .
- (iii)  $\langle \nabla \omega(x) - \nabla \omega(y), x - y \rangle \geq \alpha|x - y|^2$  for all  $x, y \in X$  and  $\lambda \in ]0, 1[$ .

*Proof.* “(i) $\Rightarrow$ (ii)”: Rewrite (i) as

$$\omega(y + \lambda(x - y)) \leq \lambda\omega(x) + (1-\lambda)\omega(y) - \frac{\alpha}{2}\lambda(1-\lambda)|x - y|^2. \quad (2.6)$$

Hence

$$\frac{\omega(y + \lambda(x - y)) - \omega(y)}{\lambda} \leq \omega(x) - \omega(y) - \frac{\alpha}{2}(1-\lambda)|x - y|^2.$$

Taking the limit as  $\lambda \rightarrow 0^+$  and using the assumption that  $\omega$  is differentiable we see that

$$\langle \nabla \omega(y), x - y \rangle \leq \omega(x) - \omega(y) - \frac{\alpha}{2}|x - y|^2.$$

Hence (ii) holds.

“(ii) $\Rightarrow$ (i)”. Suppose that (ii) holds for all  $x, y \in X$ . Let  $\lambda \in ]0, 1[$ . Set  $z = \lambda x + (1 - \lambda)y \in X$ . Applying (ii) to  $x$  and  $z$  yields

$$\omega(z) \leq \omega(x) - \langle \nabla \omega(z), x - z \rangle - \frac{\alpha}{2} |x - z|^2. \quad (2.7)$$

Similarly, applying (ii) to  $y$  and  $z$  yields

$$\omega(z) \leq \omega(y) - \langle \nabla \omega(z), y - z \rangle - \frac{\alpha}{2} |y - z|^2. \quad (2.8)$$

Multiplying (2.7) by  $\lambda$  and (2.8) by  $(1 - \lambda)$ , and adding we get

$$\begin{aligned} \omega(z) &\leq \lambda \omega(x) + (1 - \lambda) \omega(y) - \lambda \langle \nabla \omega(z), x - z \rangle - (1 - \lambda) \langle \nabla \omega(z), y - z \rangle \\ &\quad - \frac{\alpha}{2} \left( \lambda |x - z|^2 + (1 - \lambda) |y - z|^2 \right). \end{aligned}$$

Notice that  $x - z = (1 - \lambda)(x - y)$  and  $y - z = \lambda(y - x)$ . Thus, substituting in the last inequality we get

$$\begin{aligned} \omega(z) &\leq \lambda \omega(x) + (1 - \lambda) \omega(y) - \lambda \langle \nabla \omega(z), (1 - \lambda)(x - y) \rangle - (1 - \lambda) \langle \nabla \omega(z), \lambda(y - x) \rangle \\ &\quad - \frac{\alpha}{2} [\lambda (1 - \lambda)^2 |x - y|^2 + (1 - \lambda) \lambda^2 |x - y|^2] \\ &= \lambda \omega(x) + (1 - \lambda) \omega(y) - \lambda (1 - \lambda) \langle \nabla \omega(z), x - y \rangle + \lambda (1 - \lambda) \langle \nabla \omega(z), x - y \rangle \\ &\quad - \frac{\alpha}{2} \lambda (1 - \lambda) \left( (1 - \lambda) |x - y|^2 + \lambda |x - y|^2 \right) \\ &= \lambda \omega(x) + (1 - \lambda) \omega(y) - \frac{\alpha}{2} \lambda (1 - \lambda) |x - y|^2. \end{aligned}$$

Substituting for  $z = \lambda x + (1 - \lambda)y$  gives (i).

“(ii) $\Rightarrow$ (iii)”. Suppose that (ii) holds  $\forall x, y \in X$ . Then we have

$$\omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle \geq \frac{\alpha}{2} |x - y|^2, \quad (2.9)$$

$$\omega(y) - \omega(x) + \langle \nabla \omega(x), x - y \rangle \geq \frac{\alpha}{2} |x - y|^2. \quad (2.10)$$

Adding (2.9) and (2.10) we get (iii).

“(iii) $\Rightarrow$ (ii)”. By the fundamental theorem of calculus we have for  $t \in ]0, 1[$

$$\omega(x) - \omega(y) = \int_0^1 \langle \nabla \omega(y + t(x - y)), x - y \rangle dt.$$

Subtracting  $\langle \nabla \omega(y), x - y \rangle$ , noting that  $\int_0^1 \langle \nabla \omega(y), x - y \rangle dt = \langle \nabla \omega(y), x - y \rangle$  and using (iii) we get

$$\begin{aligned} \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle &= \int_0^1 \langle \nabla \omega(y + t(x - y)) - \nabla \omega(y), x - y \rangle dt \\ &= \int_0^1 \frac{1}{t} \langle \nabla \omega(y + t(x - y)) - \nabla \omega(y), t(x - y) \rangle dt \\ &\geq \int_0^1 \frac{1}{t} \alpha |t(x - y)|^2 dt \\ &= \alpha |x - y|^2 \int_0^1 \frac{1}{t} t^2 dt \\ &= \frac{\alpha}{2} |x - y|^2, \end{aligned}$$

which completes the proof.  $\square$

Following [3], we give the definition of the Bregman diameter of an arbitrary set  $X$ .

**Definition 2.13.** Let  $\omega : O \rightarrow \mathbb{R}$  be a convex differentiable function. Let  $X$  be a nonempty subset of  $O$ . The Bregman diameter of the set  $X$  is defined as

$$\Theta = \sup\{D_\omega(u, v) : u, v \in X\}. \quad (2.11)$$

In the following lemma we prove that, if  $\omega$  is differentiable and strongly convex, then the Bregman diameter is finite for every compact subset of  $\mathbb{R}^m$ .

**Lemma 2.14.** Let  $\omega : O \rightarrow \mathbb{R}$  be a differentiable convex function. Let  $X$  be a nonempty compact subset of  $O$ . Then  $D_\omega$  is bounded on  $X \times X$ . Consequently, the Bregman diameter of the set  $X$  is finite.

*Proof.* Since  $\omega$  is convex and differentiable, therefore  $\omega$  is continuously differentiable on  $O$  [21, Corollary 25.5.1]. Thus,  $\omega$  and  $\nabla\omega$  are continuous on  $X$ , and therefore  $D_\omega$  is continuous on  $X \times X$ . Now,  $X \times X$  is a nonempty compact subset of  $\mathbb{R}^m \times \mathbb{R}^m$ , and therefore  $D_\omega$  is bounded on  $X \times X$  and the Bregman diameter of the set  $X$  is finite.  $\square$

### 3 Functional Constraints and Assumptions

Recall that we are interested in the general convex problem of the form

$$(P) : \min \{f(x) : g(x) \leq 0, x \in X\}. \quad (3.1)$$

In the sequel, we shall consider the following assumptions on  $f$ ,  $g$  and  $X$ .

**A1**  $f : O \rightarrow \mathbb{R}$  and  $g : O \rightarrow \mathbb{R}$  are continuous convex functions.

**A2**  $X$  is a nonempty compact convex subset of  $O$ , and  $X$  is not a singleton.

**A3** We have access to lower- $\mathcal{C}^2$  representations (see Theorem 2.5) of  $f$  and  $g$  on some open subset  $O'$  of  $O$  such that  $X \subseteq O'$  and

$$f = \max_{t \in T_f} f_t \quad \text{and} \quad g = \max_{t \in T_g} g_t \quad \text{on } O'.$$

**A4** The set of optimal solutions of problem  $(P)$  is nonempty.

**Remark 3.1.** Under Assumption **A1**, the functions  $f$  and  $g$  are lower- $\mathcal{C}^2$  functions on  $O$  (by Lemma 2.2). Assumption **A3** provides the stronger statement that we have access to lower- $\mathcal{C}^2$  representations of these functions.

**Lemma 3.2.** Suppose that Assumptions **A1** and **A2** hold. Then

$$L_f := \sup \|\partial f(X)\| < +\infty \quad \text{and} \quad L_g := \sup \|\partial g(X)\| < +\infty. \quad (3.2)$$

*Proof.* Combine Remark 3.1, Assumption **A2**, and Lemma 2.6(ii).  $\square$



In the following Theorem, we give an error bound for the approximate subgradient.

**Theorem 3.3.** *Suppose that **A1**, **A2**, **A3**, and **A4** hold. Let  $Y = (y_0, y_1, \dots, y_m) \in \mathbb{R}^{m \times (m+1)}$  be a poised tuple of sample points centred at  $y_0 \in X$  with respect to  $f$ . Set  $\Delta = \max_{1 \leq i \leq m} |y_i - y_0|$ . Suppose that  $B(y_0; \Delta) \subseteq X$ . Let  $y \in B(y_0; \Delta)$ . Let  $(t_1, \dots, t_r) \in A(y)^r$ , and  $\lambda \in S_r$ , where  $r \in \mathbb{N}$ . Define  $V = V(y) := \sum_{i=1}^r \lambda_i \nabla F_{t_i}(y)$ . Then there exists  $v \in \partial f(y)$  such that the following error bound holds:*

$$|V - v| \leq K_f(1 + \sqrt{m}|\hat{L}^{-1}|/2) \Delta,$$

where  $K_f$  is as in Lemma 2.7, and  $\hat{L} = \hat{L}(Y)$  is as defined in Theorem 2.9.

*Proof.* By assumption  $V = \sum_{i=1}^r \lambda_i \nabla F_{t_i}(y)$ . Lemma 2.4 implies that  $v = v(y) := \sum_{i=1}^r \lambda_i \nabla f_{t_i}(y) \in \partial f(y)$ . Using the triangle inequality, the error bound given in Theorem 2.9 (applied to  $O'$  instead of  $O$ ) and Lemma 2.7, we have

$$\begin{aligned} |V - v| &= \left| \sum_{i=1}^r \lambda_i (\nabla F_{t_i}(y) - \nabla f_{t_i}(y)) \right| \leq \sum_{i=1}^r \lambda_i |\nabla F_{t_i}(y) - \nabla f_{t_i}(y)| \\ &\leq \sum_{i=1}^r \lambda_i K_f(1 + \sqrt{m}|\hat{L}^{-1}|/2) = K_f(1 + \sqrt{m}|\hat{L}^{-1}|/2) \Delta, \end{aligned}$$

as claimed.  $\square$

Our next corollary relates Theorem 3.3 to the algorithm presented later. Let us note that the function  $E$  in Corollary 3.4 is the same as the one used in the algorithm. We also note that, although in Corollary 3.4 we provide the error bound for the approximate gradient function in a general format, in practice we shall use  $x = y_0$ .

**Corollary 3.4.** *Suppose that **A1**, **A2**, **A3** and **A4** hold. Let  $Y = (y_0, y_1, \dots, y_m)$  be a poised tuple of sample points centered at  $y_0 \in X$  with respect to  $f$ . Set  $\Delta = \max_{1 \leq i \leq m} |y_i - y_0|$  and suppose that  $B(y_0; \Delta) \subseteq X$ . For every  $x \in B(y_0; \Delta)$ , let  $(t_1, \dots, t_{r(x)}) \in A_f(x)^{r(x)}$ ,  $\lambda \in S_{r(x)}$ ,  $(\bar{t}_1, \dots, \bar{t}_{\bar{r}(x)}) \in A_g(x)^{\bar{r}(x)}$ ,  $\bar{\lambda} \in S_{\bar{r}(x)}$ ,*

$$\begin{aligned} v_f(x) &= \sum_{i=1}^{r(x)} \lambda_i \nabla f_{t_i}(x) \in \partial f(x), & V_f(x) &= \sum_{i=1}^{r(x)} \lambda_i \nabla F_{t_i}(x), \\ v_g(x) &= \sum_{i=1}^{\bar{r}(x)} \bar{\lambda}_i \nabla g_{\bar{t}_i}(x) \in \partial g(x), & V_g(x) &= \sum_{i=1}^{\bar{r}(x)} \bar{\lambda}_i \nabla G_{\bar{t}_i}(x), \end{aligned}$$

and

$$e(x) := \begin{cases} v_f(x), & \text{if } g(x) \leq \varepsilon, \\ v_g(x), & \text{otherwise,} \end{cases} \quad (3.3)$$

and

$$E(x) := \begin{cases} V_f(x), & \text{if } g(x) \leq \varepsilon \\ V_g(x), & \text{otherwise.} \end{cases} \quad (3.4)$$

Then:

(i) The following error bound holds

$$|e(x) - E(x)| \leq \kappa \Delta, \text{ for all } x \in B(y_0; \Delta), \quad (3.5)$$

where  $\kappa = \max\{K_f, K_g\}(1 + \sqrt{m}|\hat{L}^{-1}|/2)$ ,  $K_f$  is defined as in Lemma 2.7 and  $K_g$  is obtained by replacing  $f$  by  $g$  in Lemma 2.7, and  $\hat{L}$  is as defined in Theorem 2.9.

(ii) The function  $E$  induced by (3.4) satisfies

$$|E(x)| \leq \max\{L_f, L_g\} + \kappa \Delta, \text{ for all } x \in B(y_0; \Delta), \quad (3.6)$$

where  $L_f$  and  $L_g$  are defined as in Lemma 3.2.

*Proof.* (i): Use (3.3) and (3.4), and apply Theorem 3.3 to  $f$  and  $g$ . (ii): Let  $x \in X$ . Using the triangle inequality, (3.2), and (3.5) we have  $|E(x)| \leq |e(x)| + |e(x) - E(x)| \leq \max\{L_f, L_g\} + \kappa \Delta$ .  $\square$

## 4 Algorithm and Discussion

In this section we introduce the Derivative-Free  $\varepsilon$ -CoMirror algorithm and present a convergence analysis.

### The Derivative-Free $\varepsilon$ -CoMirror algorithm (DFO $_{\varepsilon}$ CM)

#### **Initialization** Input

- $x_0 \in X$ ,
- $M \in \mathbb{R}_{++}$ .

#### **General step** for every $k \in \{1, 2, \dots\}$

- Select

$$0 < \Delta_k \leq \frac{1}{\sqrt{k+1}}. \quad (4.1)$$

- Select a poised tuple  $Y_k = (y_0, y_1, \dots, y_m)$  centred at  $y_0$  with respect to  $f$  such that the set  $\{y_0, y_1, \dots, y_m\} \subseteq B(x_k, \Delta_k)$ ,  $x_k = y_0$  and  $|\hat{L}_k^{-1}| \leq M$ , where  $\hat{L}_k = \hat{L}(Y_k)$  is as defined in Theorem 2.9.
- Set

$$x_{k+1} = \operatorname{argmin}_{x \in X} \{ \langle t_k E_k - \nabla \omega(x_k), x \rangle + \omega(x) \}, \quad (4.2)$$

where

$$E_k := \begin{cases} V_f(x_k), & \text{if } g(x_k) \leq \varepsilon; \\ V_g(x_k), & \text{otherwise,} \end{cases} \quad (4.3)$$

$$t_k = \frac{\sqrt{\Theta \alpha}}{|E_k| \sqrt{k}}, \quad (4.4)$$

and where  $\alpha > 0$  is the strong convexity parameter of the strongly convex function  $\omega: O \rightarrow \mathbb{R}$ ,  $\Theta$  is the corresponding Bregman diameter of the set  $X$ , and  $V_f$  and  $V_g$  are defined as in Corollary 3.4.

**Remark 4.1.**

- (i) In generating the points of the tuple  $Y_k \subseteq \mathbb{R}^{m \times (m+1)}$  we need to check that  $|\hat{L}_k^{-1}| \leq M$ . If this inequality fails, then we resample. It is always possible to generate the tuple  $Y_k$  for all  $k \in \mathbb{N}$  provided that  $M$  is set to be sufficiently large [25]. For a detailed discussion on how to choose  $M$  we refer the reader to [9].
- (ii) The poised tuple  $Y_k = (y_0, y_1, \dots, y_m)$  must satisfy  $\max_{i \in \{1, \dots, m\}} |y_i - x_k| \leq \Delta_k$  to guarantee that the error bound in Theorem 3.3 still holds true. This does not create a conflict (i) because by the definition of the matrix  $\hat{L}$  in (2.4), the value of  $|\hat{L}^{-1}|$  remains unchanged under scaling or shifting.
- (iii) The update of  $x_k$  in (4.2) is well defined, since that the function  $\langle t_k E_k - \nabla \omega(x_k), \cdot \rangle + \omega$  is strongly convex and differentiable over  $X$ , and therefore it has a unique minimizer over  $X$ .
- (iv) The step length  $t_k$  is well defined for all  $k \in \{1, 2, \dots\}$  except when  $E_k = 0$  in which case either we have a local minimum, or we change the search radius  $\Delta_k$  to get a better approximation of the gradients. Moreover, the Bregman diameter  $\Theta$  is finite by Lemma 2.14. Finally, by Lemma 2.12 (ii), we have that  $D_\omega(x, y) \geq \frac{\alpha}{2} |x - y|^2$ , and therefore, since  $X$  is not a singleton, the Bregman diameter  $\Theta$  is strictly positive.
- (v) In general, the Bregman diameter  $\Theta$  is not easy to calculate. However, if the set  $X$  is simple and the function  $\omega$  is separable, calculating  $\Theta$  becomes simpler. For example, if  $X = [\alpha_1, \beta_1] \times \dots \times [\alpha_m, \beta_m]$  and  $\omega(x) = \sum_{i=1}^m \omega_i(x_i)$ , then  $\Theta = \sum_{i=1}^m D_{\omega_i}(\alpha_i, \beta_i)$ .

## 4.1 Convergence Analysis

We devote this subsection to study the convergence of the algorithm. Lemma 4.2 and its proof are only a minor adaptation of [3, Lemma 2.2]. For the sake of completeness, we include the adapted proof.

**Lemma 4.2.** *Let  $(x_k)_{k \in \mathbb{N}}$  be the sequence generated by DFO<sub>ε</sub>CM. Let  $i < j$  be two strictly positive integers. Then for all  $k \in \{1, 2, \dots\}$*

$$\sum_{k=i}^j t_k \langle E_k, x_k - u \rangle \leq \Theta + \frac{1}{2\alpha} \sum_{k=i}^j t_k^2 |E_k|^2, \quad (4.5)$$

for every  $u \in X$ .

*Proof.* By the optimality condition in (4.2) we have

$$\langle t_k E_k - \nabla \omega(x_k) + \nabla \omega(x_{k+1}), u - x_{k+1} \rangle \geq 0 \text{ for every } u \in X.$$

Hence,

$$t_k \langle E_k, u - x_{k+1} \rangle \geq \langle \nabla \omega(x_k) - \nabla \omega(x_{k+1}), u - x_{k+1} \rangle \text{ for every } u \in X. \quad (4.6)$$

The three-point property of the Bregman distance [7, Lemma 3.1] tells us

$$D_\omega(u, x_{k+1}) - D_\omega(u, x_k) + D_\omega(x_{k+1}, x_k) = \langle \nabla \omega(x_k) - \nabla \omega(x_{k+1}), u - x_{k+1} \rangle. \quad (4.7)$$

Combining (4.6) and (4.7) yields

$$t_k \langle E_k, u - x_{k+1} \rangle \geq D_\omega(u, x_{k+1}) - D_\omega(u, x_k) + D_\omega(x_{k+1}, x_k).$$

That is

$$t_k \langle E_k, x_{k+1} - u \rangle \leq D_\omega(u, x_k) - D_\omega(x_{k+1}, x_k) - D_\omega(u, x_{k+1}).$$

Adding  $t_k \langle E_k, x_k - x_{k+1} \rangle$  to both sides of the above inequality and using Lemma 2.12 (ii) and the Cauchy-Schwarz inequality we get

$$\begin{aligned} t_k \langle E_k, x_k - u \rangle &\leq D_\omega(u, x_k) - D_\omega(u, x_{k+1}) - D_\omega(x_{k+1}, x_k) + t_k \langle E_k, x_k - x_{k+1} \rangle \\ &\leq D_\omega(u, x_k) - D_\omega(u, x_{k+1}) - \frac{\alpha}{2} |x_k - x_{k+1}|^2 + t_k |E_k| |x_k - x_{k+1}|. \end{aligned}$$

Notice that,  $t_k |E_k| |x_k - x_{k+1}| - \frac{\alpha}{2} |x_k - x_{k+1}|^2$  is a quadratic function of  $|x_k - x_{k+1}|$  that has a maximum value of  $\frac{1}{2\alpha} t_k^2 |E_k|^2$ , i.e.,  $t_k |E_k| |x_k - x_{k+1}| - \frac{\alpha}{2} |x_k - x_{k+1}|^2 \leq \frac{1}{2\alpha} t_k^2 |E_k|^2$ . This yields

$$t_k \langle E_k, x_k - u \rangle \leq D_\omega(u, x_k) - D_\omega(u, x_{k+1}) + \frac{1}{2\alpha} t_k^2 |E_k|^2.$$

Summing the last inequality over  $k \in \{i, i+1, \dots, j\}$  we obtain

$$\sum_{k=i}^j t_k \langle E_k, x_k - u \rangle \leq D_\omega(u, x_i) - D_\omega(u, x_{j+1}) + \sum_{k=i}^j \frac{1}{2\alpha} t_k^2 |E_k|^2.$$

Using the definition of  $\Theta$  we note that  $D_\omega(u, x_i) - D_\omega(u, x_{j+1}) \leq \Theta$ , from which we get (4.5).  $\square$

The following theorem presents the efficiency estimate for the Derivative-Free  $\varepsilon$ -CoMirror method. In proving Theorem 4.3 we are motivated by the techniques used in the proof of [3, Theorem 2.1]. Given  $n \in \mathbb{N}$ , we denote the set of indices of the  $\varepsilon$ -feasible solutions among the first  $n$  iterations by

$$I_n^\varepsilon = \{k \in \{1, 2, \dots, n\} : g(x_k) \leq \varepsilon\}.$$

**Theorem 4.3.** *Suppose that Assumptions A1, A2, A3 and A4 hold. Let  $\varepsilon > 0$  and let  $(x_k)_{k \in \mathbb{N}}$  be the sequence generated by DFO $_\varepsilon$ CM. Denote by  $f_{\text{opt}}$  the optimal function value of (3.1). Then for every  $n \in \{4, 5, \dots\}$*

$$\min \left\{ \min_{k \in I_n^\varepsilon} (f(x_k) - f_{\text{opt}}), \varepsilon \right\} \leq \frac{C}{\sqrt{n}},$$

where

$$\begin{aligned} C &= 2\sqrt{\frac{\Theta}{\alpha}} \max\{\kappa_1, \kappa_2\} \frac{1 + \ln(2)}{2 - \sqrt{2}} + \kappa_2 \Omega, \\ \kappa_1 &= \max\{L_f, L_g\}, \\ \kappa_2 &= K(1 + \sqrt{m}M/2), \\ \Omega &= \max_{x, y \in X} |x - y|, \end{aligned}$$

$L_f$  and  $L_g$  are as defined in (3.2),  $K$  is as defined in Corollary 3.4, and  $M > 0$  satisfies that  $|\hat{L}_k^{-1}| \leq M$  for all  $k \in \{1, 2, \dots\}$ .

*Proof.* Using assumption **A4**, suppose that  $x_{\text{opt}}$  is an optimal solution of (3.1). Fix  $n \in \{1, 2, \dots\}$ , and  $k \in \{1, 2, \dots, n\}$ . We begin by considering the following two cases:

**Case I:**  $k \in I_n^\varepsilon$ . Then  $g(x_k) \leq \varepsilon$ , and, by (4.3), (3.3), and (3.4) we have  $e_k := e(x_k) = v_f(x_k) \in \partial f(x_k)$  and  $E_k := E(x_k) = V_f(x_k)$ , and hence

$$f(x_k) \leq f(x_{\text{opt}}) + \langle e_k, x_k - x_{\text{opt}} \rangle.$$

Therefore, using Cauchy-Schwarz inequality and the error bound in equation (3.5)

$$\begin{aligned} f(x_k) &\leq f(x_{\text{opt}}) + \langle E_k, x_k - x_{\text{opt}} \rangle + \langle e_k - E_k, x_k - x_{\text{opt}} \rangle \\ &\leq f(x_{\text{opt}}) + \langle E_k, x_k - x_{\text{opt}} \rangle + |e_k - E_k| |x_k - x_{\text{opt}}| \\ &\leq f(x_{\text{opt}}) + \langle E_k, x_k - x_{\text{opt}} \rangle + \kappa_2 \Delta_k \Omega. \end{aligned}$$

Hence

$$f(x_k) - f(x_{\text{opt}}) \leq \langle E_k, x_k - x_{\text{opt}} \rangle + \kappa_2 \Delta_k \Omega. \quad (4.8)$$

**Case II:**  $k \notin I_n^\varepsilon$ . Then  $g(x_k) > \varepsilon$ . Using (4.3), (3.3), and (3.4) we have  $e_k = v_g(x_k) \in \partial g(x_k)$  and  $E_k = V_g(x_k)$ , and hence

$$g(x_k) \leq g(x_{\text{opt}}) + \langle e_k, x_k - x_{\text{opt}} \rangle.$$

Since  $g(x_{\text{opt}}) \leq 0$  we have

$$\begin{aligned} \varepsilon &< g(x_k) \\ &\leq g(x_{\text{opt}}) + \langle e_k, x_k - x_{\text{opt}} \rangle \\ &\leq \langle e_k, x_k - x_{\text{opt}} \rangle = \langle E_k, x_k - x_{\text{opt}} \rangle + \langle e_k - E_k, x_k - x_{\text{opt}} \rangle. \end{aligned}$$

Hence, using Cauchy-Schwarz inequality, the assumption that  $|\hat{L}_k^{-1}| \leq M$  for all  $k \in \{1, 2, \dots\}$ , and the error bound in equation (3.5) we have

$$\begin{aligned} \varepsilon &\leq \langle E_k, x_k - x_{\text{opt}} \rangle + |e_k - E_k| |x_k - x_{\text{opt}}| \\ &\leq \langle E_k, x_k - x_{\text{opt}} \rangle + \kappa_2 \Delta_k \Omega. \end{aligned} \quad (4.9)$$

By combining Case I and Case II, we have

$$\langle E_k, x_k - x_{\text{opt}} \rangle + \kappa_2 \Delta_k \Omega \geq \begin{cases} f(x_k) - f(x_{\text{opt}}), & \text{if } k \in I_n^\varepsilon, \\ \varepsilon, & \text{if } k \notin I_n^\varepsilon. \end{cases} \quad (4.10)$$

Using (4.10) we have for all  $1 \leq l \leq n$ , with  $\Delta_l \leq 1/\sqrt{l+1}$

$$\min\{\min_{k \in I_n^\varepsilon} (f(x_k) - f(x_{\text{opt}})), \varepsilon\} \leq \langle E_l, x_l - x_{\text{opt}} \rangle + \kappa_2 \Delta_l \Omega.$$

Let  $n_0 \in \{1, 2, \dots, n\}$ , then using (4.1)

$$\begin{aligned} \min\left\{\min_{k \in I_n^\varepsilon} (f(x_k) - f(x_{\text{opt}})), \varepsilon\right\} &\leq \min_{n_0 \leq l \leq n} (\langle E_l, x_l - x_{\text{opt}} \rangle + \kappa_2 \Delta_l \Omega) \\ &\leq \min_{n_0 \leq l \leq n} \left( \langle E_l, x_l - x_{\text{opt}} \rangle + \kappa_2 \Omega \max_{n_0 \leq l \leq n} \Delta_l \right) \\ &\leq \min_{n_0 \leq l \leq n} (\langle E_l, x_l - x_{\text{opt}} \rangle) + \frac{\kappa_2 \Omega}{\sqrt{n_0 + 1}}. \end{aligned} \quad (4.11)$$

Substituting  $u = x_{\text{opt}}$ ,  $i = n_0$ ,  $j = n$  in Lemma 4.2 we see that

$$\sum_{k=n_0}^n t_k \langle E_k, x_k - x_{\text{opt}} \rangle \leq \Theta + \frac{1}{2\alpha} \sum_{k=n_0}^n t_k^2 |E_k|^2. \quad (4.12)$$

On the other hand, since  $X$  is not a singleton, we have  $t_k > 0$  for every  $k \in \{1, 2, \dots, n\}$ , and thus

$$\sum_{k=n_0}^n t_k \langle E_k, x_k - x_{\text{opt}} \rangle \geq \left( \min_{n_0 \leq k \leq n} \langle E_k, x_k - x_{\text{opt}} \rangle \right) \sum_{k=n_0}^n t_k. \quad (4.13)$$

Combining (4.12) and (4.13) yields

$$\min_{n_0 \leq k \leq n} \langle E_k, x_k - x_{\text{opt}} \rangle \leq \frac{\Theta + \frac{1}{2\alpha} \sum_{k=n_0}^n t_k^2 |E_k|^2}{\sum_{k=n_0}^n t_k}. \quad (4.14)$$

Using (4.4), we have

$$\sum_{k=n_0}^n t_k^2 |E_k|^2 = \Theta \alpha \sum_{k=n_0}^n \frac{1}{k}, \quad (4.15)$$

and

$$\sum_{k=n_0}^n t_k = \sqrt{\Theta \alpha} \sum_{k=n_0}^n \frac{1}{|E_k| \sqrt{k}}. \quad (4.16)$$

We recall that  $|\hat{L}_k^{-1}| \leq M$  for all  $k \in \{1, 2, \dots\}$ ,  $\kappa_1 = \max \{L_f, L_g\}$  and  $\kappa_2 = K(1 + \sqrt{m}M/2)$ . Now, for every  $k \in \{1, 2, \dots\}$  using Corollary 3.4 and (4.1) we have

$$\begin{aligned} |E_k| \sqrt{k} &\leq (\kappa_1 + \kappa_2 \Delta_k) \sqrt{k} \leq \kappa_1 \sqrt{k} + \kappa_2 \frac{\sqrt{k}}{\sqrt{k+1}} \\ &\leq \kappa_1 \sqrt{k} + \kappa_2 \leq \max \{ \kappa_1, \kappa_2 \} (\sqrt{k} + 1) \\ &\leq 2 \max \{ \kappa_1, \kappa_2 \} \sqrt{k}. \end{aligned} \quad (4.17)$$

Using (4.16) and (4.17) we get

$$\sum_{k=n_0}^n t_k \geq \frac{\sqrt{\Theta \alpha}}{2 \max \{ \kappa_1, \kappa_2 \}} \sum_{k=n_0}^n \frac{1}{\sqrt{k}}, \quad (4.18)$$

Using equations (4.15) and (4.18), inequality (4.14) becomes

$$\min_{n_0 \leq l \leq n} \langle E_l, x_l - x_{\text{opt}} \rangle \leq \frac{2 \Theta \max \{ \kappa_1, \kappa_2 \} \left( 1 + \frac{1}{2} \sum_{k=n_0}^n \frac{1}{k} \right)}{\sqrt{\Theta \alpha} \sum_{k=n_0}^n \frac{1}{\sqrt{k}}}. \quad (4.19)$$

Now, set  $n_0 = \lfloor n/2 \rfloor$ . On the one hand, using (4.19) and Lemma A.1 we get

$$\min_{n_0 \leq l \leq n} \langle E_l, x_l - x_{\text{opt}} \rangle \leq \frac{C_1}{\sqrt{n}}, \quad (4.20)$$

where  $C_1 = 2\sqrt{\frac{\Theta}{\alpha}} \max\{\kappa_1, \kappa_2\} \frac{1 + \ln(2)}{2 - \sqrt{2}}$ . On the other hand, using the fact that  $\lfloor n/2 \rfloor + 1 > n/2$  we have

$$\frac{\kappa_2 \Omega}{\sqrt{n_0 + 1}} = \frac{\kappa_2 \Omega}{\sqrt{\lfloor n/2 \rfloor + 1}} \leq \frac{C_2}{\sqrt{n}}, \quad (4.21)$$

where  $C_2 = \sqrt{2} \kappa_2 \Omega$ . Using (4.20) and (4.21) we deduce that

$$\min\{\min_{k \in I_n^\varepsilon} f(x_k) - f(x_{\text{opt}}), \varepsilon\} \leq \frac{C_1 + C_2}{\sqrt{n}} = \frac{C}{\sqrt{n}}, \quad (4.22)$$

which completes the proof.  $\square$

## 5 Numerical Results

In this section we provide some numerical results of the  $\text{DFO}_\varepsilon\text{CM}$  algorithm. The  $\text{DFO}_\varepsilon\text{CM}$  algorithm was implemented in MATLAB. To begin we examine three academic test problems from [10, 11]. We then apply the  $\text{DFO}_\varepsilon\text{CM}$  algorithm to a simulation test problem from [16].

### 5.1 Academic Test Problems

We first consider three academic test problems from [10, 11]. In working with these problems, we rewrite the constraint functions as a single constraint via a max function. For example, in Test Problem 1 the constraint functions are rewritten as  $g(x_1, x_2) = \max_{1 \leq i \leq 3} g_i(x)$ , where  $g_1(x_1, x_2) = -x_1$ ,  $g_2(x_1, x_2) = x_1 - 1$  and  $g_3(x_1, x_2) = x_2$ .

(i) Test Problem 1

$$\begin{aligned} (x \in \mathbb{R}^2) \quad & \text{Minimize} \quad -x_1 - 2x_2 \\ & \text{subject to} \quad 0 \leq x_1 \leq 1 \\ & \quad \quad \quad x_2 \leq 0. \end{aligned}$$

(ii) Test Problem 2

$$\begin{aligned} (x \in \mathbb{R}^2) \quad & \text{Minimize} \quad 6x_1^2 + x_2^2 - 60x_1 - 8x_2 + 166 \\ & \text{subject to} \quad 0 \leq x_1 \leq 10, \\ & \quad \quad \quad 0 \leq x_2 \leq 10, \\ & \quad \quad \quad x_1 + x_2 - x_1x_2 \geq 0, \\ & \quad \quad \quad x_1 + x_2 - 3 \geq 0. \end{aligned}$$

(iii) Test Problem 3

$$\begin{aligned}
(x \in \mathbb{R}^2) \quad & \text{Minimize } 7x_1^2 + 3x_2^2 - 84x_1 - 34x_2 + 300 \\
& \text{subject to } 0 \leq x_1 \leq 10, \\
& \quad 0 \leq x_2 \leq 10, \\
& \quad x_1x_2 - 1 \geq 0, \\
& \quad 9 - x_1^2 - x_2^2 \geq 0.
\end{aligned}$$

**Remark 5.1.** In [10] and [11], the authors mention that their algorithms could not find an optimal solution to Test Problem 3. This is due to them incorrectly stating that the optimal value is  $-97.30952$ . The correct optimal value is  $f_{\text{opt}} \approx 84.6710$ , which we demonstrate below.

Define  $f$ ,  $g_1$ , and  $g_2$  as follows,

$$\begin{aligned}
f(x_1, x_2) &= 7x_1^2 + 3x_2^2 - 84x_1 - 34x_2 + 300, \\
g_1(x_1, x_2) &= 1 - x_1x_2 \leq 0, \text{ and} \\
g_2(x_1, x_2) &= x_1^2 + x_2^2 - 9 \leq 0.
\end{aligned}$$

Notice that  $f(x_1, x_2) = 7(x_1 - 6)^2 + 3(x_2 - \frac{17}{3})^2 - \frac{145}{3}$ , so  $f$  is strictly convex.

The constraint set  $\{(x_1, x_2) \in \mathbb{R}^2: 0 \leq x_1 \leq 10, 0 \leq x_2 \leq 10, g_1(x_1, x_2) \leq 0 \text{ and } g_2(x_1, x_2) \leq 0\}$  is also convex. Let  $a$  be the positive real root of  $p(x) = 16x^4 - 336x^3 + 1909x^2 + 3024x - 15876$ . Then at  $x_1 = a$ , and  $x_2 = \frac{8}{357}a^3 - \frac{4}{17}a^2 + \frac{145}{714}a + \frac{36}{17}$ , with  $\lambda = -1 - \frac{1909}{378}a + \frac{8}{9}a^2 - \frac{8}{189}a^3$  we have  $1 - x_1x_2 < 0$ ,  $x_1^2 + x_2^2 = 9$  and  $\nabla f(x_1, x_2) = \lambda \nabla g_2(x_1, x_2)$ ; that is first order optimality holds. As the objective function and constraint set are convex, this implies optimality. The corresponding optimal value is  $f_{\text{opt}} \approx 84.6710$ . Approximate values of  $(x_1, x_2) = (2.6390, 1.4267)$  and  $\lambda \approx -8.9150$ .

We test DFO<sub>ε</sub>CM on each of these three test problems using two options for creating the Bregman distance. In the results of these test problems we shall use  $\omega_1 = \frac{1}{2}|\cdot|^2$ , and  $\omega_2(x)$  to denote the (negative) entropy  $\sum_{i=1}^m (x_i) \ln(x_i)$ . In Table 1 we compare our results of the first three test problems to the results obtained by the Pattern Search method and Simplex Search method introduced in [10]. Note that, although in test problems 2 and 3 the constraint functions are non convex, the generated constraint set is convex. This is not covered by Theorem 4.3, however; the DFO<sub>ε</sub>CM still gives a good fit.

Examining Table 1, we note that DFO<sub>ε</sub>CM outperformed both the Pattern Search and Simplex Search algorithms on Test Problems 2 and 3. On Test Problem 1, DFO<sub>ε</sub>CM did not preform as well, but still required noticeably less function evaluations than the Pattern Search and Simplex Search methods.

## 5.2 Simulation Test Problem

In this section we test the algorithm on 12-dimensional simulated maximization problem given in [16]. We used the same starting points given in [16]:  $x_0 = (1, 0, \dots, 0)$  and  $\bar{x}_0 = (2, 0.5, \dots, 0.5)$  are vectors in  $\mathbb{R}^{12}$ . The results are reported in Table 2. We compare our results to the results obtained from the Direct Pattern Search Method (DPS) and the Direct Random Search Method with Simulated Annealing (DRS+SA) in [16]. As the constraint set for this problem is a system of



Table 1: Comparing results for Test Problems 1, 2, and 3.

Test Problem	Results	DFO CoMirror		Pattern Search	Simplex Search
		$\omega_1(x)$	$\omega_2(x)$	Algorithm[10]	Algorithm[10]
1	Function value	-0.9542	-0.9645	-1	-1
	$f$ evaluations	78	99	195	158
	$g$ evaluations	162	141	157	129
2	Function value	7.5587	7.5580	7.625	7.625
	$f$ evaluations	78	81	138	146
	$g$ evaluations	122	111	138	118
3	Function value	84.7096	84.7108	85.6610	85.6200
	$f$ evaluations	78	75	154	198
	$g$ evaluations	122	125	154	153

linear inequalities, the methods used in [16] used exact gradients when dealing with constraints. Objective function evaluations are provided via deterministic simulation.

The results in [16] report that, using 3000 function calls, the DPS gives an optimal value of 0.8327 with  $x_0$  as starting point and an optimal value of 0.1747 with  $\bar{x}_0$  as starting point. Whereas, using 3000 function calls, the heuristic DRS+SA gives an optimal value of 0.9628 with  $x_0$  as starting point and an optimal value of 0.9671 with  $\bar{x}_0$  as starting point.

Table 2: Results of DFO CoMirror algorithm

f calls	Starting point $x_0$			Starting point $\bar{x}_0$		
	$\varepsilon = 0.01$	$\varepsilon = 0.005$	$\varepsilon = 0.001$	$\varepsilon = 0.01$	$\varepsilon = 0.005$	$\varepsilon = 0.001$
100	0.7329	0	0	0.8875	0	0.8968
500	0.9400	0.9387	0.9342	0.9220	0.9210	0.8332
1000	0.9452	0.9514	0.9447	0.9277	0.9256	0.9334
3000	0.9547	0.9551	0.9546	0.9500	0.9467	0.9538

In Table 2 we see that with 500 function calls,  $\text{DFO}_\varepsilon\text{CM}$  is able to achieve a significantly better fit than the DPS. While the fit for  $\text{DFO}_\varepsilon\text{CM}$  never quite achieves the quality of the DRS+SA method, it comes quite close after 3000 function calls. This difference could be explained by the fact that the DRS+SA method employs heuristics to break free of local minimizers.

## 6 Conclusion

In this paper we developed the convergence analysis required to generate a derivative-free comirror algorithm,  $\text{DFO}_\varepsilon\text{CM}$ . Furthermore, we provided some numerical results from the implementation of the algorithm in MATLAB. One natural line of future research is to adapt the algorithm to deal with the problem

$$(P1) : \min\{f(x) : g(x) \leq 0\}, \quad (6.1)$$

i.e.,  $X = \mathbb{R}^m$ , and to prove convergence. Another line of future research is examining the convergence in the case where  $g$  is not necessarily convex, but the constraint set remains convex. Results

from test problems 2 and 3 suggest that this is possible.

## A Appendix

**Lemma A.1.** *For any integer  $n \in \{4, 5, \dots\}$  the following inequalities hold true*

$$\sum_{k=\lfloor n/2 \rfloor}^n \frac{1}{k} \leq 2\ln(2), \quad (\text{A.1})$$

$$\sum_{k=\lfloor n/2 \rfloor}^n \frac{1}{\sqrt{k}} \geq (2 - \sqrt{2})\sqrt{n}. \quad (\text{A.2})$$

*Proof.* To see inequality (A.1), notice

$$\begin{aligned} \sum_{k=\lfloor n/2 \rfloor}^n \frac{1}{k} &\leq \sum_{k=\lfloor n/2 \rfloor - 1}^{n-1} \int_k^{k+1} \frac{1}{x} dx \\ &= \int_{\lfloor n/2 \rfloor - 1}^n \frac{1}{x} dx \\ &= \ln\left(\frac{n}{\lfloor n/2 \rfloor - 1}\right). \end{aligned} \quad (\text{A.3})$$

We now consider two cases ( $n$  is even and  $n$  is odd). Case I: suppose  $n = 2m$  with  $m \in \{1, 2, \dots\}$ . Then

$$\frac{n}{\lfloor n/2 \rfloor - 1} \leq 4 \iff \frac{2m}{m-1} \leq 4 \iff n = 2m \geq 4. \quad (\text{A.4})$$

Case II: suppose  $n = 2m + 1$  with  $m \in \{1, 2, \dots\}$ . Then

$$\frac{n}{\lfloor n/2 \rfloor - 1} \leq 4 \iff \frac{2m+1}{m-1} \leq 4 \iff n = 2m+1 \geq 7. \quad (\text{A.5})$$

Moreover, for  $n = 5$  direct computation shows that  $\frac{n}{\lfloor n/2 \rfloor - 1} \leq 4$ , which together with (A.3), (A.4) and (A.5) proves the first inequality for all  $n \in \{2, 3, \dots\}$ .

Finally,

$$\begin{aligned} \sum_{k=\lfloor \frac{n}{2} \rfloor}^n \frac{1}{\sqrt{k}} &= \sum_{k=\lfloor \frac{n}{2} \rfloor}^n \frac{1}{\sqrt{k}}(k+1-k) \geq \sum_{k=\lfloor \frac{n}{2} \rfloor}^n \int_k^{k+1} \frac{1}{\sqrt{x}} dx = \int_{\lfloor \frac{n}{2} \rfloor}^{n+1} \frac{1}{\sqrt{x}} dx \\ &\geq \int_{\frac{n}{2}}^n \frac{1}{\sqrt{x}} dx = (2 - \sqrt{2})\sqrt{n}. \end{aligned}$$

which proves inequality (A.2) □

## Acknowledgments

HHB was partially supported by the Natural Sciences and Engineering Research Council of Canada and by the Canada Research Chair Program. WLH was partially supported by the Natural Sciences and Engineering Research Council of Canada and UBC Internal Research Funding. WMM was partially supported by the Natural Sciences and Engineering Research Council of Canada and UBC Internal Research Funding.

## References

- [1] M.A. Abramson, C. Audet, and J.E. Dennis, Jr. Filter pattern search algorithms for mixed variable constrained optimization problems. *Pac. J. Optim.*, 3(3):477–500, 2007.
- [2] C. Audet and J. E. Dennis, Jr. Mesh adaptive direct search algorithms for constrained optimization. *SIAM J. Optim.*, 17(1):188–217 (electronic), 2006.
- [3] A. Beck, A. Ben-Tal, N. Guttman-Beck, and L. Tetruashvili. The CoMirror algorithm for solving nonsmooth constrained convex problems. *Oper. Res. Lett.*, 38(6):493–498, 2010.
- [4] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003.
- [5] A. Ben-Tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM J. Optim.*, 12(1):79–108 (electronic), 2001.
- [6] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comp. Math. Math. Phys.*, 7:200–217, 1967.
- [7] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.*, 3(3):538–543, 1993.
- [8] A.R. Conn, K. Scheinberg, and L.N. Vicente. *Introduction to Derivative-Free Optimization*, SIAM, 2009.
- [9] A.R. Conn, K. Scheinberg, and L.N. Vicente. Geometry of interpolation sets in derivative free optimization. *Math. Program. (Ser. B)*, 111(1-2):141–172, 2008.
- [10] A. Correia, J. Matias, P. Mestre, and C. Serôdio. Derivative-free optimization and filter methods to solve nonlinear constrained problems. *Int. J. Comput. Math.*, 86(10-11):1841–1851, 2009.
- [11] A. Correia, J. Matias, P. Mestre, and C. Serôdio. Direct-search penalty/barrier methods. In *Proceedings of the World Congress on Engineering, London, U.K.*, volume III, 2010.
- [12] Z. Denkowski, S. Migórski, and N.S. Papageorgiou. *An Introduction to Nonlinear Analysis: Theory*. Kluwer, Boston, MA, 2003.

- [13] D.W. Dreisigmeyer. Direct search algorithms over riemannian manifolds. Los Alamos Technical Report LA-UR-06-7416, 2006.
- [14] D.W. Dreisigmeyer. Equality constraints, riemannian manifolds and direct search methods,. Los Alamos Technical Report LA-UR-06-7406, 2006.
- [15] N. Ghosh and W.W. Hager. A derivative-free bracketing scheme for univariate minimization. *Comput. Math. Appl.*, 20(2):23–34, 1990.
- [16] W.L. Hare. Using derivative free optimization for constrained parameter selection in a home and community care forecasting model. In *International Perspectives on Operations Research and Health Care, Proceedings of the 34th Meeting of the EURO Working Group on Operational Research Applied to Health Sciences*, pages 61–73, 2010.
- [17] S. Lucidi, M. Sciandrone, and P. Tseng. Objective-derivative-free methods for constrained optimization. *Math. Program. (Ser. A)*, 92:37–59, 2002.
- [18] R. Mifflin, A superlinearly convergent algorithm for minimization without evaluating derivatives. *Math. Program.*, 9:100–117, 1975.
- [19] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, New York, 1983.
- [20] M.J.D. Powell. UOBYQA: unconstrained optimization by quadratic approximation. *Math. Program. (Ser. B)*, 92:555–582, 2002.
- [21] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- [22] R.T. Rockafellar and R.J-B Wets. *Variational Analysis*, Springer, Berlin, 1998.
- [23] F. Vanden Berghen. *CONDOR: A Constrained, Non-Linear, Derivative-Free Parallel Optimizer for Continuous, High Computing Load, Noisy Objective Functions*. PhD thesis, Université Libre de Bruxelles, Belgium, 2004.
- [24] F. Vanden Berghen and H. Bersini. CONDOR, a new parallel, constrained extension of Powell’s UOBYQA algorithm: experimental results and comparison with the DFO algorithm. *J. Comput. Appl. Math.*, 181(1):157–175, 2005.
- [25] M. Wschebor. Smoothed analysis of  $\kappa(A)$ . *J. Complexity*, 20(1):97–107, 2004.
- [26] C. Zălinescu. *Convex analysis in general vector spaces*. World Scientific Publishing, River Edge, NJ, 2002.